ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 5,Sep. 2025,pp. 1283－1293
DOI:10. 16337/j. 1004-9037. 2025. 05. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于CNN-Transformer混合架构的AI生成图像鲁棒检测方法

康馨元[1]，李　帆[2]，赵　慧[1,4,5]，王保栋[3]，李　鑫[4,5]

(1. 济南大学信息科学与工程学院山东省泛在智能计算重点实验室(筹)，济南 250022;2. 北京埃斯顿医疗科技有限公司,北京 102200;3. 青岛大学计算机科学技术学院，青岛 266071;4. 山东省计算中心(国家超级计算济南中心)算力互联网与信息安全教育部重点实验室,济南 250353;5. 山东省基础科学研究中心(计算机科学)齐鲁工业大学(山东省科学院)山东省工业网络和信息系统安全重点实验室，济南 250353)

摘　要：深度生成模型的快速发展使得合成图像的逼真度不断提高,从图像生成到人脸篡改,各类生成技术已经深入人们的日常生活,图像真实性问题引起关注。此外,主流的图像分类模型主要在风格丰富多变的自然场景数据集上进行预训练,而单一提示词虽能生成大量的数据,但是存在明显的同质化问题,影响了学习难度的均衡性,从而使得传统的图像二分类训练方法在生成图像检测任务上存在泛化能力不足的问题。针对此类问题,本文提出了一种难易样本不均衡下的检测方法,无需修改现有分类模型,通过生成数据的自我增强方式,建立了一种有效的数据增强范式,扩充生成数据的多样性,从而平衡模型的学习难度。同时,在难易样本中利用修正的类交叉熵损失进行敏感惩罚。本文所提方法在2023年11月山东省人工智能学会举办的计算机视觉应用挑战赛(真假图片识别赛)中取得了最好的结果。

关键词：深度学习;图像分类;数据增强;真假图像识别;类别不均衡问题

中图分类号：TP391　　**文献标志码：**A

Robust Detection Method for AI-Generated Images Based on CNN-Transformer Hybrid Architecture

KANG Xinyuan[1], LI Fan[2], ZHAO Hui[1,4,5], WANG Baodong[3], LI Xin[4,5]

(1. Shandong Provincial Key Laboratory of Ubiquitous Intelligent Computing, School of Information Science and Engineering, University of Jinan, Jinan 250022, China; 2. Beijing Estun Medical Technology Co. Ltd., Beijing 102200, China; 3. College of Computer Science and Technology, Qingdao University, Qingdao 266071, China; 4. Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinnan 250353, China; 5.Qilu University of Technology (Shandong Academy of Sciences), Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Jinan 250353, China)

Abstract: With the rapid development of deep generative models, the realism of synthetic images has been continuously improving, and various generative technologies have been deeply integrated into people's

基金项目：国家自然科学基金(62103165);算力互联网与信息安全教育部重点实验室开放课题(2023ZD038);山东省自然科学基金(ZR2022ZD01)。
收稿日期：2024-02-24;**修订日期：**2024-07-20

daily life, from image generation to face manipulation, which brings attention to the authenticity of images. In addition, mainstream image classification models are mainly pre-trained on natural scene datasets with rich and varied styles, while a single prompt can generate a large amount of data, but there is an obvious homogeneity problem, which affects the imbalance of learning difficulty, thus making the traditional image binary classification training method in the generated image detection task have insufficient generalization ability. To address such issues, we propose a detection method under the difficulty and easy sample imbalance, which does not need to modify the existing classification model, and establishes an effective data augmentation paradigm by generating data self-enhancement to expand the diversity of generated data, thereby balancing the learning difficulty of the model. At the same time, we use the corrected class cross-entropy loss for sensitive punishment in difficult and easy samples. Finally, the proposed method achieves the best results in the computer vision application challenge: Real and fake image recognition competition held by the artificial intelligence society of shandong province in November 2023.

Key words: deep learning; image classification; data enhancement; real and fake image recognition; class imbalance problem

引　言

从2022年底OpenAI的ChatGPT[1]横空出世到Meta的LLaMa[2]模型开源，生成式人工智能（Artificial intelligence generated content，AIGC）在全球范围内的奇点时刻似乎愈来愈近。从DELL-E到Stable diffusion，这些强大的语言模型和图像生成模型的出现，标志着AIGC技术正经历一场革命性的进步。过去一年，基于AI图像合成和操纵技术的研究工作大量涌现，生成图像的质量和逼真度得以显著提升，很容易欺骗普通人的判断，这加剧了人们对利用虚假信息进行网络欺诈和传播的担忧[3-4]。围绕人工智能在网络安全和信息真实性方面的应用，开展自动鉴别真假图像的研究迫在眉睫，这将是保障网络环境健康的重要一环。此外，考虑到生成模型前所未有的能力及其潜在的滥用风险，设计自动识别AI图像的检测器至关重要。在这背景下，山东省人工智能学会举办的“计算机视觉应用挑战赛：真假图片识别”为算法设计提供了难得的实践平台。

相较于使用人工设计和提取特征的传统机器学习方法，深度学习在图像处理方面具有独特的优势。深度学习方法可以自动学习图像的高级特征表示，具有强大的拟合能力。同时，还可以利用大规模数据集上的预训练模型，进一步提升性能。近年来，Transformer架构因擅长建模远程依赖关系在图像分类任务中展现出优势[5-10]。然而，与基于卷积神经网络（Convolutional neural networks，CNN）的模型相比，使用一维绝对位置编码的视觉Transformer会存在较弱的局部归纳和内在多尺度表达能力，这对图像理解形成了限制。为了改善这一缺陷，针对挑战赛本文利用了一种混合的CNN-Transformer网络，通过加入了局部感知模块并借鉴了ResNet架构的分层设计，提高了局部特征提取与多尺度表示能力，形成一个融合局部与全局的高效识别框架。

在挑战赛的数据集中，由于AI生成的数据采用了相同或相似的提示词来生成，这导致了合成数据中存在明显的同质化问题。具体来说，这部分生成的图像往往在风格、构图等方面高度相似，这与真实世界的图像复杂多样性形成了反差。同时，这种数据同质化问题将严重影响模型的评估方法和泛化性能，因为模型可能只适配特定模式的训练数据，而无法很好地推广到测试集上。为了解决这个问题，本文采用了多种数据增广技术，有效增加了训练数据的多样性。这些增广方式可以模拟图像中可能出现

的变化,使模型对不同情况都具有一定的认知能力。另外,本文还通过采样和人工检查构建类别均衡的验证集,进而实现更可靠的模型评估。这些策略的引入,显著提升了模型的鲁棒性和泛化能力,是缓解数据同质化问题的有效手段。在损失函数上,传统交叉熵损失会促使模型仅关注正确类别的概率最大化,而忽略其他类别概率分布的建模。考虑到两类图像特征较为相近的情况,这种过度自信的问题会加剧。为了缓解这一问题,本文设计了一种惩罚交叉熵损失,可以更严格地惩罚错误预测,同时也提高模型对高置信预测的置信度。

综上所述,针对这项真假图像识别任务,本文试图结合当前数据集的特定问题,设计一套简单而有效的训练流程,在挑战赛提供的测试数据上对本文提出的方法进行评估,在排行榜中取得了最优性能。

1 数据集

本文使用的数据集来源于山东省人工智能学会主办的“计算机视觉应用挑战赛:真假图片识别”[11]赛道。比赛的目标是训练一个模型,通过分析图像的质量、构图和细节等方面的特征,判断未知图像是真实场景还是AI生成图像。如图1所示,AI生成的这些图像虽然用肉眼不易区分真假,但在细节上仍存在一些差异。比如生成图像中的物体边缘不够清晰自然,色彩过度饱和,场景构图不合逻辑等。此外,该竞赛数据集还具有以下特点:

(1) 数据集规模大。比赛提供的训练数据包含16 000张大小均为3像素×256像素×256像素的彩色图像,其中包括8 000张真实图像和8 000张AI生成的图像,这为模型提供了充足的训练数据。

(2) 图像质量高。数据集中的图像具有较高的视觉质量,真实图像真实地反映了现实世界中的场景,而AI生成的图像与真实图像相似度高。

(3) 数据集种类多样性丰富。数据集涵盖了各种场景,包括自然风光、人物肖像、动物、建筑、食物和艺术画等,这使得模型能够学习到各种场景下的图像特征。

(4) 数据集难易适中。数据集中的图像具有不同的难度,既有容易识别的图像,也有难以识别的图像,这使得模型能够在不同的难度下进行训练和评估。

(5) 数据同质化。由于AI生成数据中使用相同或相似的提示词来生成,导致合成图像类型存在明显的同质化,这也是该数据集的一大特点和难点。

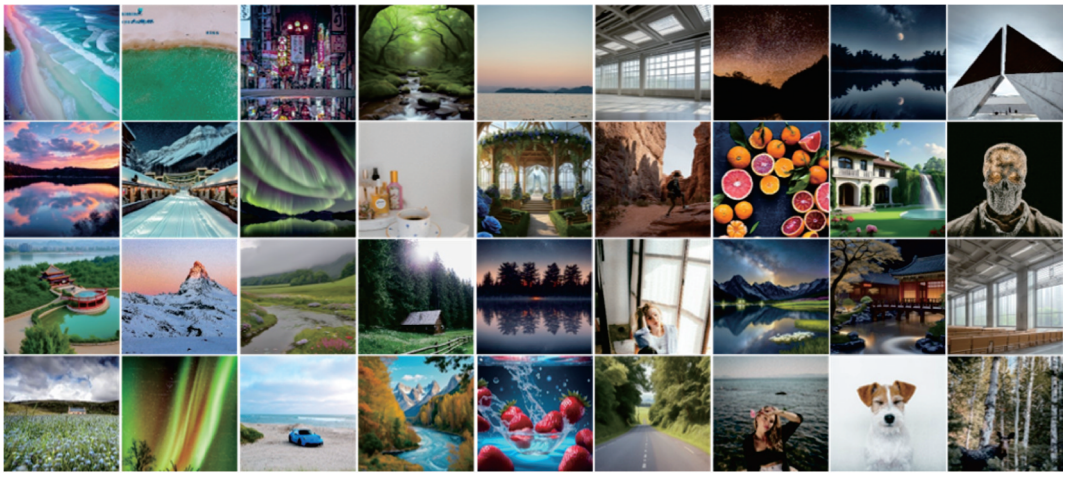

图1 AI生成图像与相机拍摄图像

Fig.1 AI-generated images and images captured by camera

2 本文方法

2.1 模型架构

本文采用一种基于CNN-Transformer的混合架构——CMT(CNNs meet transformers)[12]。如图2所示,CMT充分利用了Transformer捕捉远程依赖和CNN对局部信息建模优势。同时,由于ViT(Vision Transformer)和DeiT(Data-efficient image Transformers)等纯Transformer架构直接将输入图像分割为不重叠的图像块,这种处理方式导致每个图像块内的结构信息无法被线性投影层良好地建模。图中:Conv是卷积,stride是步长,DW-Conv是深度可分离卷积,GELU(Gaussian error linear unit)是一种高斯误差线性单元激活函数,BN是批量归一化,$H \times W$表示当前阶段输入特征图的分辨率,C为特征图的通道维度。

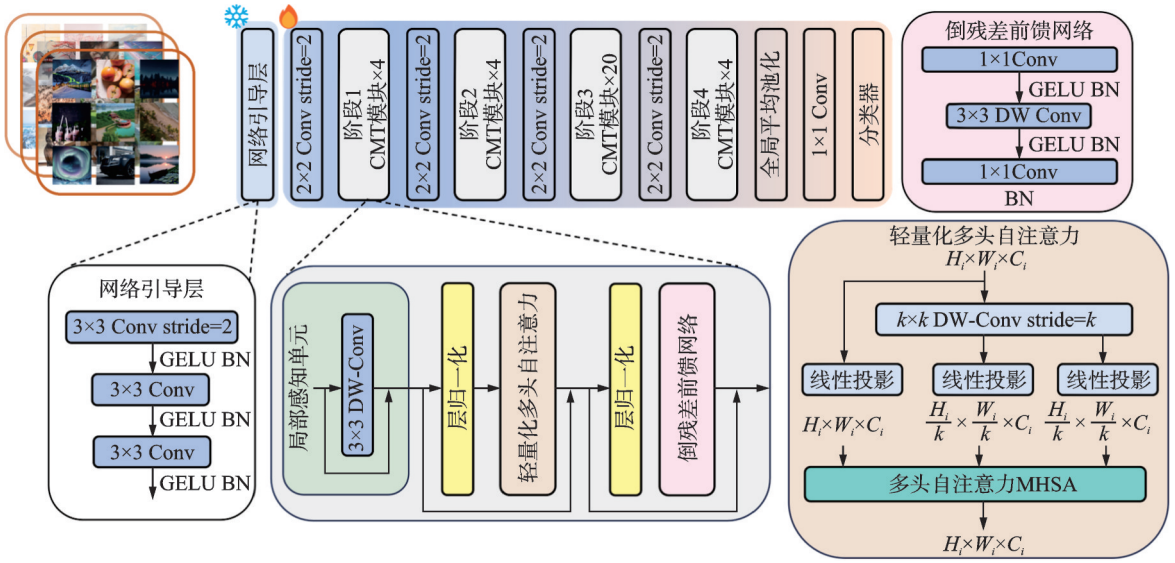

图2 本文模型架构

Fig.2 Overall architecture of the proposed model

为了解决这个问题,所用模型首先利用由3层卷积组成的网络引导层对整张输入图像进行细粒度特征提取,从而获得保留局部结构信息的特征图。随后再将特征图输入到CMT模块中进行后续的表示学习。具体来说,CMT模块是传统Transformer模块的变体,利用深度卷积对局部信息进行增强,与ViT($H/16 \times W/16$)相比,CMT第一阶段生成的特征($H/4 \times W/4$)可以保持更高的分辨率,使模型可以充分利用图像的细节信息。模型整体采用了类ResNet的分阶段架构设计,通过使用滑动步长为2的卷积层进行每个阶段前的特征下采样操作,实现逐步降低图像分辨率并增加通道维度的效果。此外,这种分阶段设计有利于提取多尺度特征,不仅增强了模型对细节和局部定位的敏感性,还具备对全局信息和语义的建模能力。同时,通过特征下采样操作,有效降低了计算成本,使其更适应分类任务的需求。

CMT模块主要由局部感知单元(Local perception unit,LPU)、轻量化多头自注意力(Lightweight multi-head self-attention,LMHSA)和倒残差前馈网络(Inverted residual feed-forward network,IRFFN)构成,通过在每个阶段堆叠不同数量的CMT块,实现特征转换和聚合。具体来说,使用残差结构组成的LPU替代Transformer中的绝对位置编码,从而捕获视觉任务中平移不变性的特点,并建立图像中的局部位置关系,其定义为

$$\mathrm{LPU}(\boldsymbol{X}) = \mathrm{DW\ Conv}(\boldsymbol{X}) + \boldsymbol{X} \tag{1}$$

式中:$\boldsymbol{X} \in \mathrm{R}^{H \times W \times C}$,DW Conv(·)为深度卷积操作。

为了降低多头自注意力的计算开销,本文所用模型构建了LMHSA,利用卷积核大小为$k\times k$和滑动步长为k的深度卷积来减少自注意力计算中的键$\boldsymbol{K}'$(其中$\boldsymbol{K}'=\mathrm{DWConv}(\boldsymbol{K})\in\mathbf{R}^{n/k^2\times d_k}$)和值$\boldsymbol{V}'$(其中$\boldsymbol{V}'=\mathrm{DW\ Conv}(\boldsymbol{V})\in\mathbf{R}^{n/k^2\times d_v}$)的空间大小,并加入一个随机初始化并可学习的相对位置偏差$\boldsymbol{B}\in\mathbf{R}^{n\times n/k^2}$,$n=H\times W$。LMHSA计算可定义为

$$\mathrm{LMHSA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}'^{\mathrm{T}}}{\sqrt{d_k}}+\boldsymbol{B}\right)\boldsymbol{V}' \tag{2}$$

式中:查询$\boldsymbol{Q}\in\mathbf{R}^{n\times d_k}$由$\boldsymbol{X}\in\mathbf{R}^{n\times d}$经线性变换得到,$d$、$d_k$和$d_v$分别对应输入、键(查询)和值的维度。受倒残差模块[13]启发,IRFFN由扩展层、深度卷积层和投影层组成,其中扩展层将输入特征的通道维度扩展4倍,而投影层将深度卷积层输出通道维度降低为原来的$\frac{1}{4}$,定义为

$$\mathrm{IRFFN}(\boldsymbol{X})=\mathrm{Conv}(F(\mathrm{Conv}(\boldsymbol{X}))) \tag{3}$$

$$F(\boldsymbol{X})=\mathrm{DWConv}(\boldsymbol{X})+\boldsymbol{X} \tag{4}$$

深度卷积用于提取局部信息,残差连接可以提高梯度的跨层传播能力。最后采用全局平均池化操作来替换ViT中的[CLS]分类标记,从而整合在空间维度上的特征,更充分地利用特征图所含的丰富信息,避免过于依赖于某个固定位置的特征。

2.2 数据集划分策略

导致训练好的模型在推理阶段脱节的关键原因之一就是所构建的验证集质量差。如图3所示,由于AI生成的合成数据使用相同或相似的提示词来生成多组图像,导致合成数据集中存在明显的同质化数据。尽管合成数据量与真实图像相当,但合成数据的实际类型远远少于真实图像,使得这些合成数据缺乏真实场景图像中的多样性和复杂性。这就导致模型过于适应训练数据的分布,而无法很好地推广到真实复杂的场景中。要解决这个问题,需要构建一个高质量的验证集,它能很好地反映真实场景的复杂性、多样性和分布,对模型进行全面的验证。同时还需要采用其他技术手段,如数据增广、正则化等方法来增强模型的泛化能力,避免模型过于适应特定的数据集,以便很好地迁移到真实场景。

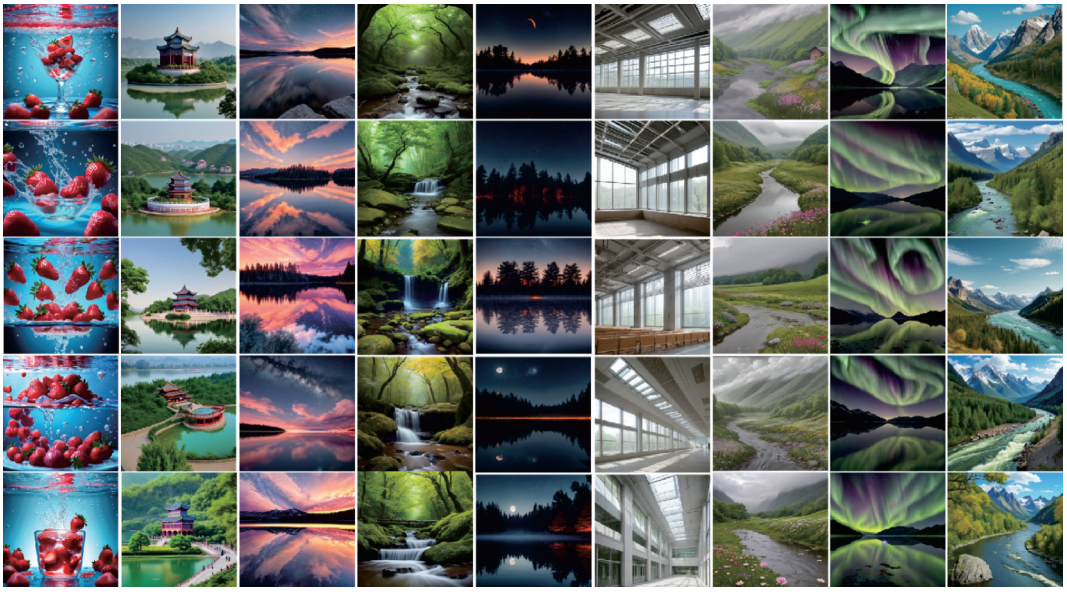

图3 相同提示词生成的同质化图像

Fig.3 Homogenized images generated from the same prompt

此外,以往常见的划分训练集和验证集的方法,如K折交叉验证或随机划分,在当前任务中会出现类别不平衡和数据泄漏问题,这是因为这类划分方式在本竞赛数据上难以保证训练集和验证集中各类的比例保持一致,容易导致某些类的数量不足,影响模型在这些类的学习效果。另外,由于生成数据集中存在的大量同质化样本,随机划分也不可避免地使训练集和验证集中存在内容高度重复的样本,这会导致验证集无法真正反映模型的泛化能力。

为了解决上述问题,本文采用了人工校正的交叉验证方式。如图4所示,首先粗略统计了数据集中由AI生成数据的类型数量,并以此数量为先验进行$K(K>25)$簇的聚类,筛选出25种类型的合成数据,将每一类型图像数据分组保存至不同的文件夹中,并利用五折交叉验证的方式分别对真实数据集和合成数据集进行划分。

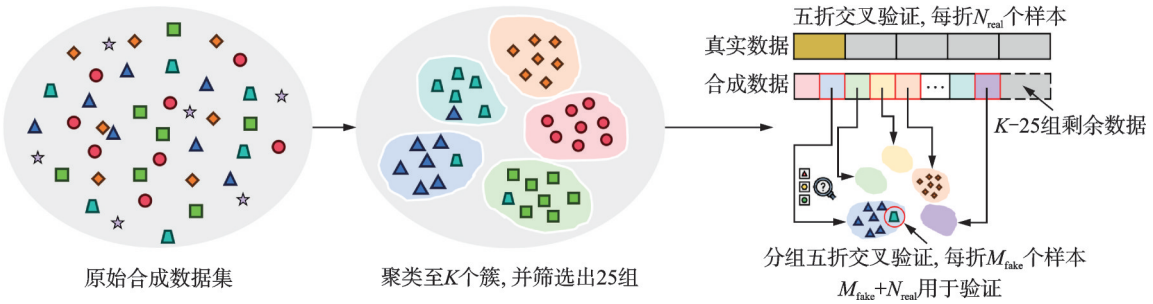

图4 人工校正的交叉验证划分方法

Fig.4 Manually-corrected cross-validation partitioning method

具体来说,对于任意一折上的训练,首先在分组后的合成数据集中任选5组共M_{fake}张合成图像,并人工筛查剩余20组中不存在所选5组类型图像;随后对于真实数据集,直接进行五折交叉验证,选择任一折的数据共N_{real}张,并将全部所选数据($M_{fake}+N_{real}$张)作为验证集。最后将两类上的剩余数据用于模型的训练。这种划分方式可以尽可能地保证训练集和验证集在类别分布和样本重复上的合理性,从而使模型验证结果更加准确可信。

2.3 数据增广策略

为了进一步扩充数据的多样性,缓解AI生成数据集中存在的同质化问题,本文加入了4类数据增广策略(图5)用于合成数据,包括随机遮挡、自身增广、硬性混叠和柔性混叠。以下是对每种策略的详细描述。

图5 数据增广方式示例

Fig.5 Illustrative examples of the data augmentation methods

随机遮挡是通过在图像上随机选择区域进行像素抹除来产生新的训练样本。这种方式让模型不过分依赖图像的局部显著性区域信息,促进模型充分利用图像中更加抽象和稳定的特征来进行分类,提高模型对部分遮挡的鲁棒性。本文使用了两种遮挡方式:

(1) 随机矩形遮挡。以一定概率根据预设超参数范围内的尺寸大小与长宽比生成矩形遮挡掩码。具体来说,设置了1个遮挡概率p_{rect},表示每张图像被遮挡的概率。对于被遮挡的图像,随机选择1个矩形区域,其尺寸大小在$[w_{\min}, w_{\max}]\times[h_{\min}, h_{\max}]$范围内,其中$w_{\min}$、$w_{\max}$、$h_{\min}$和$h_{\max}$分别为图像的最小和最大宽度和高度。然后将该矩形区域内的像素值设置为0,以模拟遮挡效果。

(2) 随机区域遮挡。将图像分为若干块区域,每个区域都以一定概率生成掩码。具体来说,将图像均匀地划分为$n\times n$个小区域,其中n为预设整数。然后,对于每个小区域,以概率p_{reigon}生成1个掩码,将该区域内的像素值设置为0,这样可以模拟图像中局部遮挡的情况。实验中将p_{rect}设置为0.3,p_{region}设置为0.2,n设置为32。

这两种遮挡方式都可以在保留图像主要内容的前提下,通过遮挡部分像素信息来增强模型的泛化能力。仅在一定概率下对图像进行遮挡,避免过度增广导致的过拟合。

自身增广是传统的包含几何变换与颜色变换等在内的增广方式,本文设计了两种自身增广策略:

(1) 图像局部增强。将待变换的图像均分为4部分,每部分按照一定概率进行水平翻转和垂直翻转,最终将4部分重新拼成原始大小的图像。具体来说,对于每一部分,以概率p_{flip}进行水平翻转,以概率p_{flip}进行垂直翻转,这样可以提供更多样化的图像特征,促使网络以更通用的方式学习局部特征。实验中发现将p_{flip}设置为0.5时可以取得较好的效果。

(2) 图像裁剪粘贴。将1张图像的任意区域进行提取并缩放,然后粘贴到图像中的随机位置。具体来说,随机选择1个矩形区域,其尺寸大小在$[w_{\text{crop}}, h_{\text{crop}}]$范围内,其中$w_{\text{crop}}$和$h_{\text{crop}}$是预设的裁剪宽度和高度。然后,将该区域内的图像进行提取,并缩放到原始图像的大小。最后,将缩放后的图像粘贴到原始图像中的随机位置,这样可以模拟图像局部遮挡的情况,同时又保留了全局内容信息。实验中发现将w_{crop}和h_{crop}设置为图像尺寸的25%时,可以取得较好的效果。

硬性混叠是利用从傅里叶空间获得的低频图像应用阈值得到的掩码,直接将2张图像进行拼接,以生成全新的图像。具体来说,先对两张图像分别进行傅里叶变换提取低频分量,然后设置1个阈值T,得到1个二值掩码。对于每个像素位置,如果该位置的低频分量大于阈值T,则将该位置的像素值设置为1,否则设置为0。然后将2张图像按该掩码进行拼接,得到混合图像,这样可以有效地融合2张图像的内容信息模拟复杂场景。实验中发现将阈值T设置为低频分量的平均值时,可以取得较好的效果。

柔性混叠是在硬性混叠的基础上设置掩膜评价权重,通过控制权重,平滑地在2张图像之间进行插值转换,得到新的混合图像。具体来说,先对2张图像分别进行傅里叶变换提取低频分量,然后设置1个掩膜评价权重α。对于每个像素位置,如果该位置的低频分量大于阈值T,则将该位置的像素值设置为α,否则设置为$1-\alpha$。然后根据该权重对2张图像进行插值,得到混合图像,这样可以得到更加平滑自然的图像融合效果。实验中发现将掩膜评价权重α设置为0.5时,可以取得较好的效果。

2.4 损失函数

传统交叉熵损失函数在训练过程中的优化目标是最大化正确类别的预测概率,而忽略了其他类别概率分布的建模。这可能导致模型在预测时过度自信,即对置信度较低的预测结果给予过高的置信度,从而引起校准问题。考虑到当前任务的特点,这一问题会进一步加剧。

为了解决上述问题,本文在损失函数上进行了改进,设计了一种惩罚交叉熵损失,它可以更严格地惩罚错误预测,同时提高模型对正确预测的置信度。具体来说,损失函数由两部分组成:传统交叉熵损

失L_{ce}和惩罚损失L_p。传统交叉熵损失L_{ce}可以表示为

$$L_{ce}(\theta)=-\frac{1}{|G|}\sum_{x_i\in G}\sum_{j=1}^{N}y_{ij}\log P(j|x_i;\theta) \tag{5}$$

式中：G表示样本集合，x_i表示第i个样本，N表示类别数量，y_{ij}表示第i个样本属于第j个类别的标签(0或1)，$P(j|x_i;\theta)$表示模型对第i个样本属于第j个类别的预测概率。

为了更严格地惩罚错误预测，引入了惩罚损失L_p。具体来说，对于每个样本x_i，计算其预测概率$P(x_i;\theta)$和真实标签y_i之间的差异，并根据差异的大小给予不同的惩罚。惩罚损失L_p可以表示为

$$L_p(\theta)=\frac{1}{|G|}\sum_{x_i\in G}\sum_{j=1}^{C}P(j|x_i;\theta)\cdot 1\{y_{ij}\neq P(j|x_i;\theta)\} \tag{6}$$

式中：$1\{y_{ij}\neq P(j|x_i;\theta)\}$表示1个指示函数，当第$i$个样本的预测概率和真实标签不同时，取值为1，否则为0。综上，最终的交叉熵损失函数可以表述为

$$L(\theta)=L_{ce}(\theta)+\lambda L_p(\theta) \tag{7}$$

式中：λ是一个权衡参数，在本文实验中设置为0.5。通过引入惩罚损失L_p，可以更严格地惩罚错误预测，同时提高模型对正确预测的置信度，这有助于提高模型的鲁棒性和泛化能力。

3 实验分析与比较

3.1 实验设置

本实验所有模型均基于Ubuntu 20.04，Conda 3.9，Pytorch 1.13.1和CUDA 11.7.1的环境实现，并在单张NVIDIA RTX A4000计算平台上进行训练、验证与测试数据推理。训练阶段，模型采用了在ImageNet预训练的CMT-B权重对除最后一层分类器外的特征提取部分进行初始化，并固定了网络引导层的参数，以充分利用在大规模数据集上训练得到的泛化特征表示，并对分类器层从头进行训练，从而适应当前任务的类别信息，此外还发现无偏差的全连接分类器表现更好。

训练阶段将所有输入图像大小均使用随机缩放裁剪方法调整至224像素×224像素，验证和推理阶段直接缩放至224像素×224像素。同时，在训练过程中应用梯度累加和混合精度技术来实现更大的批大小，以及使用早停来防止模型过拟合。表1列举了从网格搜索中得到的训练所用超参数，单折训练过程需要约17 min。

在推理阶段，机器学习竞赛的解决方案往往依赖于多样化的模型集成和测试阶段的增强技术。因此为了测试集提高分类精度，本文结合了多折上的模型版本，形成模型集成，以取得更好的泛化性能。同时，本文采用多尺度和翻转推理，对同一张图片生成5份增强结果，融合其Softmax输出获得比赛最终的提交结果。

表1 训练阶段相关超参数

Table 1 Hyper-parameters during training phase

参数类型	参数值
批大小	256
训练轮次	15
优化器	AdamW
权重衰减	5e−2
初始学习率	3e−3
学习率调整方法	余弦退火热重启
余弦退火周期	15
学习率下限	3e−6

3.2 损失实验分析

在模型训练过程中，本文同时监控训练集和验证集上的损失函数值如图6所示。从训练损失曲线可以看出，随着迭代轮数的增加，训练损失持续稳定下降，这表明模型在训练集上不断优化和学习。初期验证集损失有小幅波动，这可能是由于模型还没有完全适应数据分布造成的。在前7个epoch后，验证损失趋于稳定下降，这表明模型逐渐提取出合适的特征表示，并逐步改善了对验证集的泛化能力。稳定下降的验证损失曲线也说明模型没有出现明显的过拟合。

3.3 对比实验分析

在与现有方法的比较中,本文选择了经典的卷积神经网络和当前基于视觉Transformer的主流方法[14-22],以及竞赛排行榜前5名方法(本文方法为Top 1,竞赛排行榜第2~5名分别为Top 2~5)作为比较基准。实验结果如表2所示,本文方法在验证集和测试集的F_1-Score上分别取得了98.81%和97.84%的准确度,显著优于其他对比方法的性能。此外可以看出,与传统卷积神经网络模型相比,视觉Transformer中的空间长距离依赖建模优势具有更佳的性能表现。同时,模型的参数量也是一个重要因素。本文所用的CMT模型相比于参数量庞大的Swin Transformer V2和ConvNext V2在验证集和测试集上的F_1-score略低。但CMT采用了更为轻量化的设计,降低了计算成本。与纯Transformer结构如ViT相比,卷积的引入也弥补了Transformer对局部细节建模的不足。CMT通过这种适度的混合设计实现了较好的性能与效率权衡。

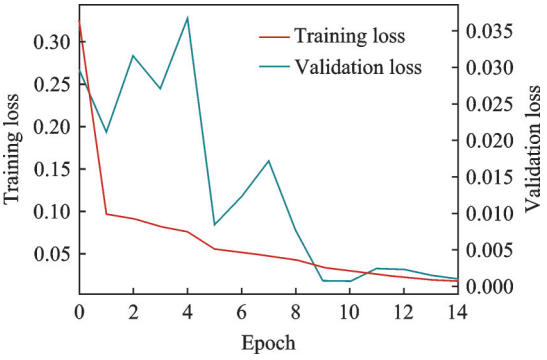

图6 训练集与验证集上的损失变化曲线

Fig.6 Loss curves for the training and validation datasets

表2 最优参数下的模型性能对比

Table 2 Comparison of model performance under optimal parameters

	模型	参数量/10^6	FLOPs/10^9	分辨率/(像素×像素)	验证集F_1-Score/%	测试集F_1-Score/%
排行榜	Top 1(本文)	**45.7**	**9.3**	**224×224**	98.81	**97.84**
	Top 2	—	—	—	—	97.18
	Top 3	—	—	—	—	96.69
	Top 4	—	—	—	—	96.46
	Top 5	—	—	—	—	96.41
主流分类模型	ResNet-152[14]	60.2	11.5	224×224	93.92	93.12
	DenseNet-169[15]	14.2	3.4	224×224	90.54	89.23
	ResNeXt-101-64x4d[16]	84	32.1	224×224	92.70	91.87
	EfficientNet-B5[17]	30.3	10.2	456×456	95.58	94.75
	EfficientNetV2-M[18]	54.1	24.58	480×480	97.04	96.62
	MaxViT-T[19]	30.9	5.56	224×224	98.77	97.69
	ViT-B/16[20]	86.5	17.56	256×256	98.52	97.29
	SwinTransV2-B[21]	87.9	20.32	256×256	99.10	98.08
	ConvNext V2-L-22k[22]	198	34.4	224×224	**99.57**	**98.34**

4 结束语

本文针对真假图像分类任务的特点,基于在ImageNet上预训练的CMT-B模型,利用其在建模长距离依赖和提取局部多尺度特征方面的优势,实现了较好的性能与资源占用的权衡。此外,针对数据集中存在较高比例的同质化数据导致的类别不平衡问题,设计了4种数据增广方式,在不改变图像类别的前提下,让有限的数据产生更丰富的特征表示效果。同时,还提出了人工筛选的数据划分方式,有针

对性地构建验证集,使得模型验证结果更加准确可靠。损失函数方面,设计了一种惩罚交叉熵损失以更严格地惩罚错误预测,并提高对正确预测的置信度。与当前同类方法相比,本文所提出的训练框架实现了更优秀的分类性能,在山东省人工智能学会举办的"计算机视觉应用挑战赛:真假图片识别"任务中取得了最好的结果,也为基于视觉的真假图像检测提供了有效的技术路线,为后续研究奠定了基础。

参考文献:

[1] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[2] TOUVRON H, LAVRIL T, IZACARD G, et al. LLAma: Open and efficient foundation language models[EB/OL]. (2023-02-05)[2024-05-02]. https://techstartups. com/wp-content/uploads/2023/02/333078981_693988129081760_4712707815225756708_n.pdf.

[3] WODAJO D, ATNAFU S, AKHTAR Z. Deepfake video detection using generative convolutional vision transformer[EB/OL]. (2023-06-13)[2024-05-02]. https://doi.org/10.48550/arXiv.2307.07036.

[4] TARIQ S, LEE S, WOO S. One detector to rule them all: Towards a general deepfake attack detection framework[C]//Proceedings of the Web Conference. Ljubljana, Slovenia: Association for Computing Machinery, 2021: 3625-3637.

[5] XU Chen, SUI Xiubao, LIU Jia, et al. Transformer in optronic neural networks for image classification[J]. Optics & Laser Technology, 2023, 165: 109627.

[6] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//Proceedings of International Conference on Machine Learning. Beijing, China: MLR, 2021: 10347-10357.

[7] GUO Menghao, LU Chengze, LIU Zhengning, et al. Visual attention network[J]. Computational Visual Media, 2023, 9(4): 733-752.

[8] PAN Zizheng, CAI Jianfei, ZHUANG Bohan. Fast vision transformers with Hilo attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 14541-14554.

[9] WU H P, XIAO B, CODELLA N, et al. CVT: Introducing convolutions to vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Canada, Montreal: IEEE, 2021: 22-31.

[10] LIU Ze, LIN Yutong, CAO Yue, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2021: 10012-10022.

[11] 山东省人工智能学会. 第十五届山东省大学生科技节第六届山东省大学生人工智能大赛暨计算机视觉应用挑战赛的通知[EB/OL]. (2023-10-28)[2024-05-05]. https://www.sdaai.org.cn/newsinfo/6503345.html. https://www.sdaai.org.cn/newsinfo/6503345.html.

Shandong Artificial Intelligence Society. Notice of the 15th Shandong University Student Science and Technology Festival and the 6th Shandong University Student Artificial Intelligence Competition and Computer Vision Application Challenge[EB/OL]. (2023-10-28)[2024-05-05]. https://www.sdaai.org.cn/newsinfo/6503345.html.

[12] GUO Jianyuan, HAN Kai, WU Han, et al. CMT: Convolutional neural networks meet vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: ArXiv E-prints, 2022: 12175-12185.

[13] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: ArXiv E-prints, 2018: 4510-4520.

[14] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: ArXiv E-prints, 2016: 770-778.

[15] GAO H, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: ArXiv E-prints, 2017: 4700-4708.

[16] XIE S N, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: ArXiv E-prints, 2017: 1492-

1500.

[17] TAN M X, LE Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 6105-6114.

[18] TAN M X, LE Q. EfficientNetV2: Smaller models and faster training[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 10096-10106.

[19] TU Z Z, TALEBI H, ZHANG H, et al. MaxViT: Multi-axis vision Transformer[C]//Proceedings of European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 459-479.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03)[2024-05-08]. https://arxiv.org/pdf/2010.11929/1000.

[21] LIU Ze, HU Han, LIN Yutong, et al. Swin Transformer V2: Scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: ArXiv E-prints, 2022: 12009-12019.

[22] WOO S, DEBNATH S, HU R H, et al. ConvNext V2: Co-designing and scaling convnets with masked autoencoders[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: ArXiv E-prints, 2023: 16133-16142.

作者简介:

康馨元(2001-),女,硕士研究生,研究方向:机器学习和储备池计算,,E-mail:kangxy2001@163.com。

李帆(1996-),男,硕士研究生,研究方向:深度学习、医学图像处理和计算机辅助诊断。

赵慧(1987-),**通信作者**,女,讲师,研究方向:复杂动态网络、多智能体,E-mail:ise_zhaohui@ujn.edu.cn。

王保栋(1995-),男,博士研究生,研究方向:计算机视觉和计算机图形学。

李鑫(1990-),男,讲师,研究方向:软件安全、网络安全和深度学习。

(编辑:刘彦东)